Popularity Predictions in Tattoo-Book Publishing

—

Dr. Anna Felicity Friedman
Director, Center for Tattoo History and Culture

Popularity and Immortality

What do authors care about most when publishing books?

- That their books will sell well
- That their books will be archived for consumption by future generations

Titles are the first thing that grab the attention of a potential reader/purchaser of a book. What if we had a calculator based on potential titles that could help predict purchase popularity as measured by archival presence in libraries?

CENTER FOR
TATTOO
HISTORY AND CULTURE

Private: The Tattoo Title Popularity Calculator

What book title might grab purchasers' attention?
Have you wondered whether a word choice between, say, "Tattoo" or "Ink" would make your book stand out more from the sea of tattooing publications that seems to increase every year?

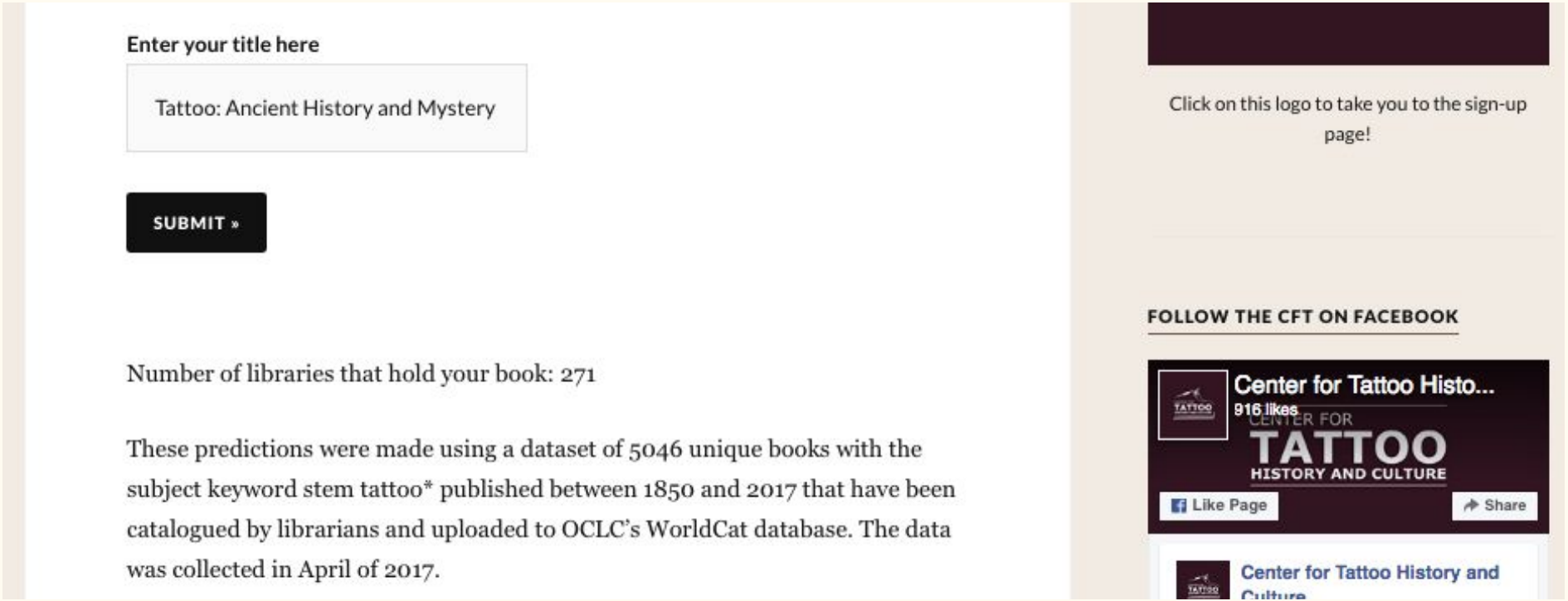
Try out different book-title options below in the CFT's Tattoo Title Popularity Calculator to find out how many libraries your book might end up in over time:

Enter your title here

SUBMIT »

SUBSCRIBE TO OUR MAILING LIST FOR NEWS AND UPDATES!

Click on this logo to take you to the sign-up page!



Users will input their proposed titles and the model will output the total count of libraries that would hold this book in their collection.

The current working model predicts this value ~6.5% better than a random guess.

Lulu Titlescorer

Book title

0 words, 0 characters. First letter:

Title is

(•) Literal ( ) Figurative Help

Title grammar type is a

Grammatically complete phrase Help

Title first word is a

Preposition, article Help

Title second word is a

Preposition/article Help

Does the title include the name of a person or place?

( ) Yes (•) No

Analyze my title!

What Do I Do?

Enter your novel title in the field at the top of the page. Use the drop-down menus to choose the variables which best describe the attributes of your title.
Click "Analyze my title!"
Score represents the percentage chance of its being a number one hit.
Results are between 9% and 83% chance of bestseller success.

E-mail this page to a friend

Learn more about Lulu

The Data

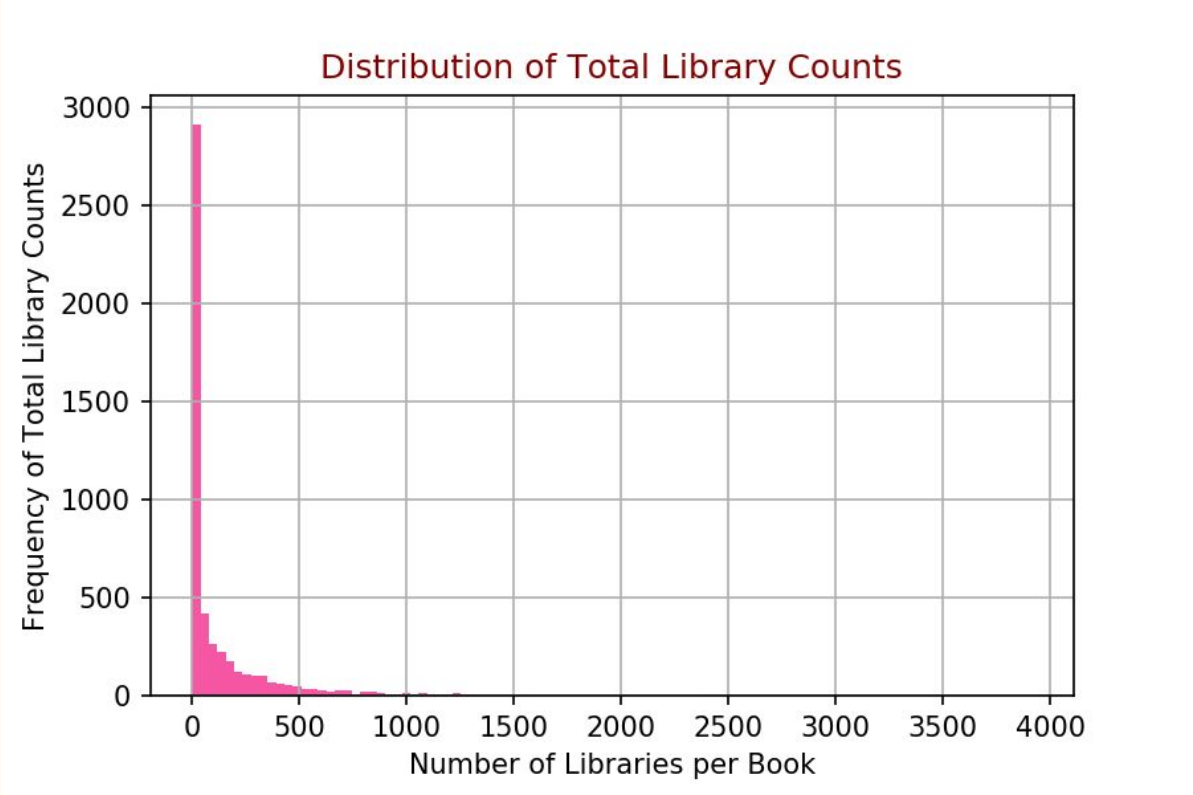
Using a combination of API queries and two different web scraping scripts, I narrowed my data to base a regression model on 5046 OCLC WorldCat library catalog records for books published 1850-2017 with the keyword stem tattoo*.

The data I pulled included:

- title, author, publication information
- summary, short title, and genre/language information (when available)
- unique OCLC record identifier
- the counts of library holdings for each edition
- ISBNs (when present)

Data Problems

- Distributions skewed heavily right and also many high-value outliers
- Approximately 15% duplicates of various sorts (entity matches due to multiple records for the same book, actual duplicates, entry mistakes, and a handful of duplicate titles that are actually different books)
- Initial data pulled using the API and web scrapers was 9391 rows
- Very noisy data: multiple foreign languages, all different book types (theses, gov't publications, ebooks, braille books)
- Narrowed to 5046 rows by limiting to printed books, English-language, with duplicates dropped




```
In [5]: data["extracted_libcount"].
```

Out[5]:

1	1178
2	312
3	196
4	127
5	92
6	81
7	76
8	64
9	55
13	49
10	48
20	43
12	41
11	38
18	35
14	35
22	28
15	27
30	27
23	27
29	25
21	24
25	23

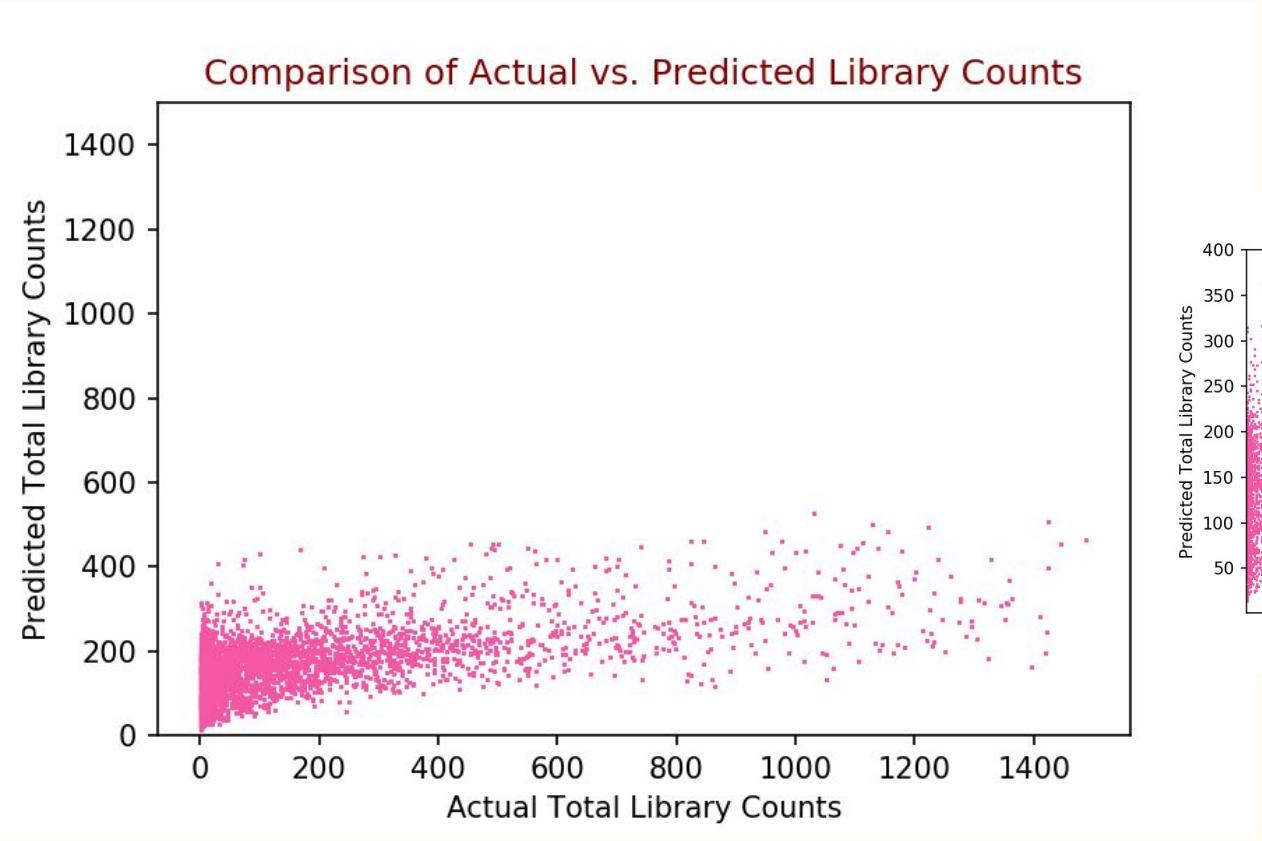
A challenge: dealing with heavily skewed data

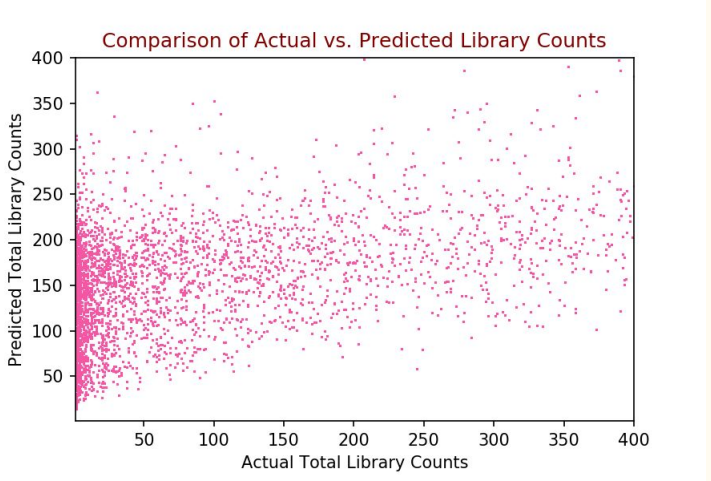
Feature Engineering

- Age (2017 minus year of publication)
- Number of words in title, title length
- Reading level (using textstat)
- Sentiment polarity and subjectivity (using textblob)
- Topic modeling (using Latent Dirichlet Allocation)
- Tf-idf word vectorization

The Model

- Sci-kit Learn's RandomForestRegressor model
- Held out a test set of one third of my data
- Test set predictions compared to the actual library count values resulted in nearly matching r-squared and explained variance scores of .063 and .064
- Mean absolute error of the test-set predictions vs. actual values was 165 (the full range of library counts in the modeling sample was 2 to 1500)
- Predicted vs. actual values on the entire data set had r-squared and explained variance rise to .26 and the mean absolute error decline to 146





The current model generally overpredicts low values and underpredicts high values

Some Predictions

1	Title	Age	Prediction
2	Ancient Ink: the historical mystery of tattoos	0	160
3	Ancient Ink: the historical mystery of tattoos	5	156
4	Ancient Ink: the historical mystery of tattoos	10	154
5	I love cats	0	95
6	Tattoos Rock	0	77
7	Tattoos Rock	2	65
8	Tattoos Rock	10	79
9	Tattoo Rock	0	114
10	Tattooing Rocks	0	113
11	Rocking Tattoos	0	66
12	The Girl with the Dragon Tattoo	0	168
13	The Girl with the Dragon Tattoo	5	180
14	Maori Moko	0	112
15	Maori Moko--Tattoos from Polynesia	0	145
16	Tattooing in Maori Culture	0	202
17	Marking Identity: Maori Tattoos and Cultural History	0	250

Conclusions

- Titles can be somewhat predictive of the popularity and immortality of a book
- A calculator such as this could be one of several tools in an author's toolbox to help them make educated decisions about what to title their book

The Future

- Try out different combination models on segments of the data, for example one that classifies the titles into one of three classes to ameliorate the overfitting problem with low values and the underfitting problem with high values

- Try a classification model rather than a regression model to predict popularity groups rather than exact numbers, for example less than 10, 10 to 100, over 100, over 1000

- Experiment with further narrowing the range of total library counts for the existing model to see if modeling on, say, 10 to 1500 works better than 2 to 1500

- Do more work with adding in certain stopwords to the model

Questions?